(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
18 December 2003 (18.12.2003)

PCT

(10) International Publication Number
WO 03/104813 A2

(51) International Patent Classification[7]: G01N 33/68

(21) International Application Number: PCT/GB03/02451

(22) International Filing Date: 6 June 2003 (06.06.2003)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

PCT/GB02/02601	7 June 2002 (07.06.2002)	GB
PCT/GB02/02778	7 June 2002 (07.06.2002)	GB
02257095.6	14 October 2002 (14.10.2002)	EP

(71) Applicants *(for all designated States except US)*: **XZILLION GMBH & CO. KG** [DE/DE]; Industriepark Hochst, Building G865, D-65926 Frankfurt am Main (DE). **PROTEOME SCIENCES PLC** [GB/GB]; Coveham House, Downside Bridge Road, Cobham, Surrey KT11 3EP (GB).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **THOMPSON, Andrew, Hugin** [GB/GB]; 30 Canterbury Street, Cambridge CB4 3QF (GB). **HAMON, Christian** [FR/DE]; Luthmerstrasse 49, D-65934 Frankfurt am Main (DE). **KUHN, Karsten** [DE/DE]; Querstrasse 29, D-44139 Dortmund (DE). **BAUER, Ute** [DE/DE]; Sandweg 129, D-60316 Frankfurt am Main (DE). **MEUMANN, Thomas** [DE/DE]; Loreleistrasse 26, D-65929 Frankfurt am Main (DE).

(74) Agents: **HILL, Christopher, Michael** et al.; Page White & Farrer, 54 Doughty Street, London WC1N 2LS (GB).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO, SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 03/104813 A2

(54) Title: CHARACTERISING POLYPEPTIDES

(57) Abstract: Provided is a method of solubilising a polypeptide, which polypeptide is insoluble or sparingly soluble in an aqueous medium, which method comprises either: (a) contacting the polypeptide with a sequence specific cleavage agent in a non-aqueous medium, to produce a product which is soluble in an aqueous medium; or (b) contacting the polypeptide with a dicarboxylic anhydride to modify one or more amino groups in the polypeptide, to produce a product which is soluble in an aqueous medium. Also provided is a relational database comprising: (a) a table of parent polypeptide sequences; and (b) a table of peptides, which peptides are expected products of putative reactions of polypeptides from the table of polypeptides; wherein each of the parent polypeptide sequences is relatable by the database to one or more peptides from the table of peptides.

CHARACTERISING POLYPEPTIDES

This invention relates to compounds and methods for characterising and/or isolating peptides from complex protein mixtures, particularly from protein aggregates that are sparingly soluble or insoluble in aqueous solutions. The procedures typically involve solubilisation of the aggregates by cleavage or chemical modification of the polypeptides followed by isolation of a subset of peptides generated by further cleavage. The isolated peptides represent the parent proteins and can be used to determine which proteins are present in a complex mixture. The invention also relates to a database facilitating determination of the parent polypeptide once the peptides have been identified.

The analysis of complex protein mixtures to determine their composition, i.e. the identities and quantities of their constituent proteins is a key part of the process of understanding the molecular basis of disease. In particular comparisons between diseased and normal tissue are essential to determine which proteins are changing in the disease process. This gives important insight into the molecular basis of disease and identifies new targets for drug discovery.

Conventionally this sort of protein analysis is carried out by performing 2-D gel electrophoresis to separate the proteins followed by mass spectrometry to identify the proteins. While the mass spectrometric identifications are robust and can be automated, 2-D gel electrophoresis is expensive and complicated to automate. As a result, new methods of analysis that can identify proteins in a cost effective and automatable process are highly sought after. One novel approach, known as MudPIT, is to digest a complex protein sample with an enzyme such as trypsin followed by analysis of the peptides by liquid chromatography mass spectrometry. The complexity of the peptide sample is reduced by extensive multi-dimensional chromatography followed by mass spectrometric analysis. This sort of process is automatable but the crude peptide digest contains many peptides for each protein and is thus very complex to analyse, requiring many fractionation steps for really complex samples increasing the labour involved. To overcome this complexity problem, the peptide analysis approach has been improved by

the use of 'sampling' procedures to isolate a subset of peptides for each protein in a mixture.

In one such peptide sampling approach, Gygi *et al.* (Nature Biotechnology 17: 994 – 999, "Quantitative analysis of complex protein mixtures using isotope-coded affinity tags" 1999) describe the use of 'isotope encoded affinity tags' for the capture of peptides from proteins, to allow protein expression analysis. In this article, the authors describe the use of a biotin linker, which is reactive to thiols, for the capture of peptides with cysteine in them. A sample of protein from one source is reacted with the biotin linker and cleaved with an endopeptidase. The biotinylated cysteine-containing peptides can then be isolated on avidinated beads for subsequent analysis by mass spectrometry. Two samples can be compared quantitatively by labelling one sample with the biotin linker and labelling the second sample with a deuterated form of the biotin linker. Each peptide in the samples is then represented as a pair of peaks in the mass spectrum where the relative peak heights indicate their relative expression levels.

Published international patent application WO 98/32876 discloses further methods of sampling peptides to identify proteins in a complex mixture. In this application, a population of proteins is profiled by isolating a single peptide from one terminus of each protein in the population. The process described comprises the steps of:

1. capturing a population of proteins onto a solid phase support by one terminus of each protein in the population;
2. cleaving the captured proteins with a sequence specific cleavage agent;
3. washing away peptides generated by the cleavage agent not retained on the solid phase support;
4. releasing the terminal peptides retained on the solid phase support; and
5. analysing the released terminal peptides, preferably identifying and quantifying each peptide in the mixture. The analysis is preferably performed by mass spectrometry.

In this application, the C-terminus is discussed as being more preferable as the terminus by which to capture a population of proteins, since the N-terminus is often blocked. In order to capture a population of proteins by the C-terminus, the C-terminal carboxyl group must be distinguished from other reactive groups on a protein and must be reacted specifically with a reagent that can effect immobilisation. In many C-terminal sequencing chemistries the C-terminal carboxyl group is activated to promote formation of an oxazolone group at the C-terminus. During the activation of the C-terminal carboxyl, side chain carboxyls are also activated, but these cannot form an oxazolone group. It has been reported that the C-terminal oxazolone is less reactive to nucleophiles under basic conditions than the activated side-chain carboxyls, offering a method of selectively capping the side chain carboxyl groups (V. L. Boyd *et al.*, Methods in Protein Structure Analysis: 109-118, Plenum Press, Edited M. Z. Atassi and E. Appella, 1995). Other more reactive side chains can be capped prior to the activation of the carboxyls using a variety of conventional reagents. In this way all reactive side chains can be capped and the C-terminus can be specifically labelled.

EP A 0 594 164 and EP B 0 333 587 describe methods of isolating a C-terminal peptide from a protein in a method to allow sequencing of the C-terminal peptide using N-terminal sequencing reagents. In this method the protein of interest is digested with an endoprotease, which cleaves at the C-terminal side of lysine residues. The resultant peptides are reacted with DITC polystyrene which reacts with all free amino groups. N-terminal amino groups that have reacted with the DITC polystyrene can be cleaved with trifluoroacetic acid (TFA) thus releasing the N-terminus of all peptides. The epsilon-amino group of lysine is not cleaved however and all non-terminal peptide are thus retained on the support and only C-terminal peptides are released. According to this patent the C-terminal peptides are recovered for micro-sequencing.

(Kaplan and Oda 1983) and DE A 4344425 (1994) describe methods of isolating an N-terminal peptide from a protein by reacting the protein with a capping reagent which will cap any free amino groups in the protein. The protein is then cleaved, and if trypsin

α-amino groups in the non-N-terminal peptides. In the first disclosure (Anal. Biochem.) the α-amino groups are reacted with dinitrofluorobenzene (DNF) which allows the non-N-terminal peptides to be captured by affinity chromatography onto a polystyrene resin while the N-terminal peptides flow through unimpeded. In DE A 4344425, the epsilon amino groups are reacted with an acylating agent prior to cleavage. After cleavage in this method, the α-amino groups on the non-N-terminal peptides are reacted with an amine reactive solid support such as diisothiocyanato glass, leaving the N-terminal peptides free in solution.

All these peptide sampling methods are useful tools to produce profiles of their parent peptides but they all ideally require soluble proteins as most of these methods require enzymatic cleavage at some stage in the sampling process. Many proteins, particularly membrane proteins, are not very soluble in aqueous media and form aggregates when they are extracted from their source tissue. They may be solubilised in non-aqueous organic solvents but most enzymes are not effective in organic solvents. Membrane proteins, however, are the receptors for most endogenous ligands and most pharmaceutical ligands and it is thus extremely important to be able to analyse these proteins.

It is an aim of the present invention to solve the problems associated with the above prior art methods. Accordingly it is an object of this invention to provide methods of solubilising protein aggregates to enable peptide sampling to take place. It is a further object of this invention to provide methods of analysing and identifying proteins in samples containing proteins that are not soluble in aqueous media, such as membrane proteins.

Thus, the present invention provides a method of solubilising a polypeptide, which polypeptide is insoluble or sparingly soluble in an aqueous medium, which method comprises either:

(a) contacting the polypeptide with a sequence specific cleavage agent in a non-aqueous medium, to produce a product which is soluble in an aqueous medium; or

(b) contacting the polypeptide with a dicarboxylic anhydride to modify one or more amino groups in the polypeptide, to produce a product which is soluble in an aqueous medium.

In the context if the present invention, the aqueous medium may be any medium comprising water, and is typically a medium in which protein samples are contained e.g. after the sample has been extracted from a subject. A non-aqueous medium is a medium which generally does not comprise water, and is a medium in which the polypeptide is at least sparingly soluble, and preferably soluble. In step (a) the cleavage agent is an agent which is soluble in the non-aqueous medium.

In the present context, the term polypeptide comprises not only polypeptides themselves, but includes also a protein, an oligopeptide or other amino acid-based molecule.

The sequence specific cleavage agent is not especially limited, provided that it is capable of reacting with the polypeptide in the medium employed. Preferably, the sequence specific cleavage agent comprises cyanogen bromide, BNPS-skatole or iodosobenzoic acid. Typically, the sequence specific cleavage agent cleaves at a methionine residue, a tryptophan residue, a cysteine residue, a threonine residue or a serine residue.

When the modification step (b) is employed, typically the dicarboxylic acid caps the one or more amino groups. Any dicarboxylic acid may be employed in this step, but preferably succinic anhydride, maleic anhydride, citraconic anhydride, dimethylmaleic anhydride and/or phthalic anhydride are employed.

Steps (a) and (b) are preferably carried out in an organic solvent, such that the proteins being modified are soluble. However, any solvent may be used, provided that the cleavage, or the modification, reaction proceeds satisfactorily. Preferred solvents include,

The present invention also provides a method for characterising a polypeptide which is insoluble or sparingly soluble in an aqueous medium, which method comprises:

(a) solubilising the polypeptide according to a method as defined above, to form a solubilised product;

(b) optionally cleaving the solubilised product to form one or more peptides;

(c) identifying one or more peptides characteristic of the polypeptide;

(d) characterising the polypeptide on the basis of the one or more identified peptides.

In this method of the invention the one or more peptides are typically isolated, e.g. by capture on a solid phase. This facilitates removal of unwanted products before characterisation is carried out. Generally the one or more peptides characteristic of the polypeptide are identified using mass spectrometry, although other characterisation methods known in the art may be employed if desired.

In one embodiment of the present invention when dicarboxylic anhydride is used in solubilising the polypeptide, it is removed before identifying the one or more peptides This typically generates unmodified peptides that are different from the unmodified proteins, as the capping changes the cleavage sites of some cleavage agents such as Trypsin.

In some preferred embodiments, the one or more peptides are separated by liquid chromatography prior to identifying the peptides.

It is also preferred that the method further comprises comparing the identified peptides with peptides in a database, in which combinations of peptides are relatable to parent polypeptides, to characterise the polypeptide. The databases that may be used in the present methods are described below.

The present invention further provides a method of producing a database for identifying a polypeptide, which method comprises:

(a) selecting one or more parent polypeptides;

(b) calculating one or more peptide sequences that would result from one or more putative reactions that each parent polypeptide could undergo;

(c) storing the calculated peptide sequences in a database such that the sequences that would result for a specific parent polypeptide undergoing its selected reaction are relatable to that parent polypeptide when undergoing that reaction.

Typically a plurality of putative reactions are selected for one or more of the parent polypeptides. Any reactions may be utilised, but typically the putative reactions are selected from the solubilising reactions and sequence specific cleavage reactions already described above. One or more of the putative reactions may be a multi-step reaction, e.g. comprising a first solubilising step and a subsequent sequence specific cleavage step. Preferably this method is carried out using a computer program in which the products of particular reactions of polypeptides are pre-programmed. The present invention also provides database obtainable by a method as defined above. The invention further provides a computer-readable storage medium comprising a database obtainable according to a method as defined above.

Further to the above, the present invention also provides a relational database comprising:

(a) a table of parent polypeptide sequences; and

(b) a table of peptides, which peptides are expected products of putative reactions of polypeptides from the table of polypeptides;

wherein each of the parent polypeptide sequences is relatable by the database to one or more peptides from the table of peptides.

Also provided is a kit for characterising a polypeptide, which kit comprises:

(a) a solubilisation agent;

(b) optionally a reagent for modifying a reactive group in a peptide or polypeptide;

(c) a sequence specific cleavage agent; and

(d) a means for selectively isolating a subset of peptides generated by the cleavage agent.

The invention will now be described in more detail by way of example only, with reference to the following specific embodiments.

In a first aspect the invention provides a method of characterising a polypeptide in a mixture containing polypeptides that are not soluble in aqueous media comprising the following steps:

- solubilising the protein mixture either:
  a. by cleavage of the sample with a sequence specific cleavage reagent that is compatible with organic solvents; or
  b. capping the free amino groups in the polypeptide with a dicarboxylic anhydride;
- isolating 1 or more characteristic peptides from the proteins in the sample;
- optionally removing the dicarboxylic anhydride if they have been used as the solubilisation reagent;
- optionally separating the sampled peptides by chromatography; and
- characterising the isolated peptides by determining their mass or sequence by mass spectrometry.

In a second aspect this invention provides a method of predicting the modified peptide products of a peptide sampling process that would be obtained from a list of known polypeptide sequences comprising the following steps:

- determining the products of the solubilisation of the polypeptide sequences, where the polypeptides are either cleaved with a sequence specific cleavage reagent that is compatible with organic solvents or amino groups in the polypeptides are reacted with a carboxylic anhydride; and
- determining the peptide products, including the expected peptide mass and sequence, of a sampling process, where the products of the solubilisation process may be treated with sequence specific cleavage reagents and/or the reactive groups

in the resultant sample peptides have been modified with selective capping reagents, followed by isolation of the selected peptides;

wherein the above determinations are calculated using a computer program.

In a third aspect this invention provides the product of a computer program on a computer-readable storage medium having stored thereon:

- a relational database comprising:
    a. a sample peptide table including a plurality of sample peptide records, each of said peptide records specifying the expected product of a peptide sampling process, in which sampling process the parent polypeptide has been cleaved and/or the reactive functionalities in the cleavage peptides have been modified with predetermined reagents; and
    b. a polypeptide sequence table including a plurality of polypeptide sequence records, each of said polypeptide sequence records specifying a parent polypeptide sequence from which the aforementioned sample peptide records have been derived;

wherein there is a many-to-many relationship between said peptide records and said polypeptide sequence item records and one polypeptide sequence item record corresponds to more than one probe record and at least one probe record corresponds to more than one polypeptide sequence item record.

In a fourth aspect the invention provides a method of identifying a polypeptide in a mixture containing polypeptides that are not soluble in aqueous media comprising the following steps:

- solubilising the protein mixture either:
    a. by cleavage of the sample with a sequence specific cleavage reagent that is compatible with organic solvents; or
    b. capping the free amino groups in the protein with a dicarboxylic anhydride;
- isolating 1 or more characteristic peptides from the proteins in the sample;

- determining mass of the isolated peptides by mass spectrometry; and
- comparing the masses of the isolated peptides with a database of known and predicted masses for proteins expected to be in the sample mixture.

In a fifth aspect this invention provides a kit comprising:

- a solubilisation reagent;
- optional reactive group modification reagents;
- a sequence specific cleavage reagent to all peptide sampling; and
- a means to selectively isolate a subset of peptides generated by the cleavage reagent.

The invention will now be described in greater detail by way of example only with reference to the following Figures.

Figure 1 shows a flow-chart outlining an algorithm to predict the sequences and masses of sampled peptides according to the methods of this invention.

Figure 2 shows an example of relational tables for storage of the data produced by the sampled peptide prediction algorithm of this invention.

Various aspects of this invention will now be discussed in greater detail.

Solubilisation Reagents

The methods of this invention comprise a method of analysing a polypeptide mixture comprising polypeptides that are not soluble in aqueous media. The solubilisation of these polypeptides is performed prior to further analysis steps and the solubilisation process modifies the parent polypeptide population.

Preferred cleavage agents are chemical reagents which are soluble in organic solvents and are volatile permitting easy removal of unreacted reagent. Appropriate chemical cleavage

reagents include cyanogen bromide which cleaves at methionine residues (Smith 1994; Smith 1997) . Under appropriate conditions this reagent will also cleave at tryptophan. A further preferred reagent is BNPS-skatole which cleaves at tryptophan residues (Crimmins, McCourt *et al.* 1990; Vestling, Kelly *et al.* 1994). Iodosobenzoic acid also cleaves at tryptophan residues (Mahoney and Hermodson 1979; Fontana, Dalzoppo *et al.* 1981; Fontana, Dalzoppo *et al.* 1983). Other chemical cleavage reagents are known including reagents such as pentafluoropropionic acid that cleave at aspartic acid residues (Tsugita, Takamoto *et al.* 1992; Tsugita, Kamo *et al.* 1998) and S-ethyltrifluorothioacetate which cleaves at threonine and serine (Kamo and Tsugita 1998). In addition reagents that cleave at cysteine (Wu and Watson 1998) are also known but reagents that cleave at methionine or tryptophan are preferred as these amino acids are comparatively rare in typical proteins. As a result, these reagents produce a relatively small number of fragments from a typical protein, meaning that the complexity of the sample is not increased greatly by the solubilisation step.

Dicarboxylic anhydrides are a second preferred choice to act as solubilisation reagent. Dicarboxylic anhydrides such as succinic anhydride, maleic anhydride, citraconic anhydride, dimethylmaleic anhydride and phthalic anhydride (described in Palacian, Gonzalez *et al.* 1990) may all be used to solubilise protein aggregates. All of these reagents react with primary amino groups in polypeptides to form amides while exposing a free carboxylic acid residue in the capping group. The negatively charged carboxylic acid groups in the capped proteins facilitate the solubilisation of protein aggregates. The capping reaction can take place in organic solvents. In addition the capping groups can be eliminated removed by raised temperature or pH allowing the unmodified peptides to be recovered (Nieto and Palacian 1983).

Sampling of peptides from a complex protein mixture

A sampling process in the context of this invention can be defined generally as a process in which a polypeptide is cleaved with a sequence specific cleavage reagent, after which specific amino acid residues are modified by specific chemical reactions and from the resultant peptide digest, peptides with specific sequence features, such as the presence of

a specific amino acid or modified amino or sequence of amino acids, are isolated for further analysis by mass spectrometry. In the context of this invention the sampling process takes place after the solubilisation process according to the methods of this invention. The aforementioned solubilisation process produces a polypeptide population that is modified with respect to the parent population and it is from this modified population that specific peptides are sampled. These result in novel peptides not produced by methods in the prior art, which have utility in the identification of the polypeptides that constitute the parent polypeptide mixture.

Reagents for the modification of amino acid side-chains

The peptide sampling methods of this invention generally require modification of at least one amino acid side-chain. Cysteine disulphide bridges are typically reduced and the free thiols are then blocked. Various methods are known in the art resulting in different mass modifications of cysteine. Since thiols are very much more reactive than the other side-chains in a protein this thiol capping step can be achieved highly selectively.

Various reducing agents have been used for disulphide bond reduction. The choice of reagent may be determined on the basis of cost, or efficiency of reaction and compatibility with the reagents used for capping the thiols (for a review on these reagents and their use see Jocelyn P.C., Methods Enzymol. **143**: 246-256, 'Chemical reduction of disulfides.' 1987).

Typical capping reagents include N-ethylmaleimide, iodoacetamide, vinylpyridine, 4-nitrostyrene, methyl vinyl sulphone or ethyl vinyl sulphone (see for example Krull L. H. & Gibbs D. E. & Friedman M., Anal. Biochem. **40**(1): 80-85, '2-Vinylquinoline, a reagent to determine protein sulfhydryl groups spectrophotometrically.' 1971; Masri M. S. & Windle J. J. & Friedman M., Biochem Biophys. Res. Commun. **47**(6): 1408-1413, 'p-Nitrostyrene: new alkylating agent for sulfhydryl groups in reduced soluble proteins and keratins.' 1972; Friedman M. & Zahnley J.C. & Wagner J.R., Anal. Biochem. **106**(1): 27-34, 'Estimation of the disulfide content of trypsin inhibitors as S-beta-(2-pyridylethyl)-L-cysteine." 1980).

Typical reducing agents include mercaptoethanol, dithiothreitol (DTT), sodium borohydride and phosphines such as tributylphosphine (see Ruegg U. T. & Rudinger J., Methods Enzymol. 47:111-116, 'Reductive cleavage of cysteine disulfides with tributylphosphine.', 1977) and tris(carboxyethyl)phosphine (Burns J.A. *et al.*, J Org Chem. 56: 2648-2650, 'Selective reduction of disulfides by tris(2-carboxyethyl)phosphine.', 1991). Mercaptoethanol and DTT may be less preferred for use with thiol reactive capping reagents as these compounds contain thiols themselves.

Amino groups are often blocked in the methods of this invention. Preferred reagents for the purposes of this invention retain a charge or ionisable functionality in the modified lysine residue. Pyridyl propenyl sulphone and other related reagents are disclosed in PCT/GB02/02601. These reagents react quite selectively with amino groups, particularly if free thiols have been capped prior to this reaction. In addition the pyridine functionality provides an ionisable group. The amino group of lysine is also retained as the alkenyl sulphone are Michael reagents producing an alkylated derivative of lysine. N-succinimidyl-2(3-pyridyl)acetate (SPA) (Cardenas, van der Heeft *et al.* 1997) is another reagent appropriate for capping amino-groups that retains and ionisable functionality in the modified lysine residues. A range of other amine reactive capping reagents appropriate for use with mass spectrometry have been developed and are reviewed in (Roth, Huang *et al.* 1998).

Terminal peptide isolation for global protein expression profiling

Isolation of N- or C-terminal peptides has been described as a method to determine a global expression profile of a protein sample. Isolation of terminal peptides ensures that at least one and only one peptide per protein is isolated thus ensuring that the complexity of the sample that is analysed does not have more components than the original sample. Reducing large polypeptides to shorter peptides makes the sample more amenable to analysis by mass spectrometry. Methods for isolating peptides from the termini of polypeptides are discussed in WO 98/32876, WO 00/20870, PCT/GB02/02778 and PCT/GB02/02601 the contents of which are incorporated herein by reference. N-terminal

peptides can be isolated, according to these disclosures by capping the amino groups followed by cleavage with sequence specific cleavage reagents such as trypsin exposing alpha-amino groups in the non-N-terminal peptide fragments which can be coupled with an amino-reactive biotin reagent such as EZ-Link™-PEO–LC– NHS-Biotin (Pierce Warriner, UK) to allow the non-N-terminal peptides to captured onto an avidinated support leaving the desired N-terminal peptides in solution. According to the methods of this invention, pre-cleavage of the parent polypeptide sample with cyanogen bromide leads to a new polypeptide population from each of which an N-terminal peptide will be isolated resulting in several peptides being analysed for each polypeptide in the parent polypeptide sample. In an embodiment of the second aspect of this invention, a protocol for the analysis of a protein sample containing polypeptides by isolating terminal peptide fragments comprises the steps of:

- pre-cleaving the polypeptides with cyanogen bromide;
- optionally reducing and capping all cysteine residues;
- blocking all free amino groups in the cyanogen bromide fragments;
- cleaving the polypeptides with a sequence specific endoprotease, such as trypsin;
- biotinylating the exposed alpha-amino groups in the non-N-terminal fragments generated by the sequence specific endoprotease;
- capturing biotinylated peptides onto an avidin derivitised solid support; and
- detecting the N-terminal peptides left in solution by LC-MS or LC-MS/MS.

In an alternative embodiment of the second aspect of this invention, a protocol for the analysis of a sample of polypeptides by isolating terminal peptide fragments from the polypeptides comprises the steps of:

- capping the free amino groups in the protein with a dicarboxylic anhydride;
- optionally reducing and capping all cysteine residues;
- cleaving the polypeptides with a first sequence specific cleavage reagent, such as cyanogen bromide;
- biotinylating the exposed alpha-amino groups in the cleavage fragments generated by the first sequence specific cleavage reagent;

- cleaving the biotinylated fragments with a sequence specific endoprotease such as trypsin;
- capturing the biotinylated N-terminal peptides onto an avidin derivitised solid support; and
- detecting the captured N-terminal peptides by LC-MS or LC-MS/MS.

Isolation of peptides containing cysteine

As discussed earlier, Gygi *et al.* (Gygi, Rist *et al.* 1999) disclose the use of 'isotope encoded affinity tags' for the capture of peptides from proteins, to allow protein expression analysis. The authors report that a large proportion of proteins (>90%) in yeast have at least one cysteine residue (on average there are ~5 cysteine residues per protein). Reduction of disulphide bonds in a protein sample and capping of free thiols with iodoacetamidylbiotin results in the labelling of all cysteine residues. The labelled proteins are then digested, with trypsin for example, and the cysteine-labelled peptides may be isolated using avidinated beads. These captured peptides can then be analysed by liquid chromatography tandem mass spectrometry (LC-MS/MS) to determine an expression profile for the protein sample. Two protein samples can be compared by labelling the cysteine residues with a different isotopically modified biotin tag. This approach is slightly more redundant than an approach based on isolating terminal peptides as, on average, more than one peptide per protein is isolated so there are more peptide species in the sample than protein species. This increase in complexity is made worse by the nature of the tags used by Gygi *et al.*

In an embodiment of the second aspect of this invention, a protocol for the analysis of a protein sample containing polypeptides with cysteine residues comprises the steps of:

- pre-cleavage of the polypeptides with cyanogen bromide;
- reducing and reacting all cysteine residues in at least one protein sample with a cysteine reactive biotin reagent, such as N-(3-maleimidopropionyl)biocytin (Fluka) or EZ-Link™ PEO-Iodoacetyl Biotin (Pierce & Warriner, UK, Ltd);
- cleaving the polypeptides with a sequence specific endoprotease;
- capturing biotinylated peptides onto an avidin derivitised solid support; and

- detecting the isolated peptides by LC-MS or LC-MS/MS.

The protein samples may be digested with the sequence specific endoprotease before or after reaction of the sample with the biotin reagent.

Isolation of glycopeptides from carbohydrate modified proteins

Carbohydrates are often present as a post-translational modification of proteins. Various affinity chromatography techniques for the isolation of these sorts of proteins are known (For a review see Gerard C., Methods Enzymol **182**: 529-539, "Purification of glycoproteins." 1990). A variety of natural protein receptors for carbohydrates are known. The members of this class of receptors, known as lectins, are highly selective for particular carbohydrate functionalities. Affinity columns derivitised with specific lectins can be used to isolate proteins with particular carbohydrate modifications, whilst affinity columns comprising a variety of different lectins could be used to isolate populations of proteins with a variety of different carbohydrate modifications. In one embodiment of the second aspect of this invention, a protocol for the analysis of a sample of proteins, which contains carbohydrate modified proteins, comprises the steps of:

- pre-cleavage of the polypeptides with cyanogen bromide or capping the free amino groups in the protein with a dicarboxylic anhydride;
- treating the sample with a sequence specific cleavage reagent such as Trypsin or Lys-C;
- passing the protein sample through affinity columns containing lectins or boronic acid derivatives to isolate only carbohydrate modified peptides; and
- detecting the isolated peptides by LC-MS or LC-MS/MS.

Note that in the protocol described above and later protocols, Lys-C would not be appropriate if solubilisation has been carried out by reaction of lysine with a dicarboxylic anhydride.

Many carbohydrates have vicinal-diol groups present, i.e. hydroxyl groups present on adjacent carbon atoms. Diol containing carbohydrates that contain vicinal diols in a 1,2-

cis-diol configuration will react with boronic acid derivatives to form cyclic esters. This reaction is favoured at basic pH but is easily reversed at acid pH. Resin immobilised derivatives of phenyl boronic acid have been used as ligands for affinity capture of proteins with cis-diol containing carbohydrates. A further protocol for sampling glycopeptides from polypeptides in a complex mixture containing carbohydrate modified polypeptides comprises the steps of:

- pre-cleavage of the polypeptides with cyanogen bromide or capping the free amino groups in the protein with a dicarboxylic anhydride;
- reacting at least one protein sample at basic pH with a boronic acid derivatised biotin reagent;
- cleaving the polypeptides with a sequence specific endoprotease;
- capturing biotinylated peptides onto an avidin derivitised solid support; and
- detecting the isolated peptides by LC-MS or LC-MS/MS.

The sample may be digested with the sequence specific endoprotease before or after reaction of the sample with the biotin reagent.

Vicinal-diols, in sialic acids for example, can also be converted into carbonyl groups by oxidative cleavage with periodate. Enzymatic oxidation of sugars containing terminal galactose or galactosamine with galactose oxidase can also convert hydroxyl groups in these sugars to carbonyl groups. Complex carbohydrates can also be treated with carbohydrate cleavage enzymes, such as neuramidase, which selectively remove specific sugar modifications leaving behind sugars, which can be oxidised. These carbonyl groups can be tagged allowing proteins bearing such modifications to be detected or isolated. Hydrazide reagents, such as Biocytin hydrazide (Pierce & Warriner Ltd, Chester, UK) will react with carbonyl groups in carbonyl-containing carbohydrate species (E.A. Bayer *et al.* , Anal. Biochem. **170**: 271 – 281, "Biocytin hydrazide – a selective label for sialic acids, galactose, and other sugars in glycoconjugates using avidin biotin technology", 1988). Alternatively a carbonyl group can be tagged with an amine modified biotin, such as Biocytin and EZ-Link™ PEO-Biotin (Pierce & Warriner Ltd, Chester, UK), using reductive alkylation (Means G.E., Methods Enzymol 47: 469-478, "Reductive alkylation

of amino groups." 1977; Rayment I., Methods Enzymol 276: 171-179, "Reductive alkylation of lysine residues to alter crystallization properties of proteins." 1997). Proteins bearing vicinal-diol containing carbohydrate modifications in a complex mixture can thus be biotinylated. Biotinylated, hence carbohydrate modified, proteins may then be isolated using an avidinated solid support.

Thus a further method of sampling peptides from proteins in a mixture to represent the proteins in the sample comprises the following steps:

- pre-cleavage of the polypeptides with cyanogen bromide or capping the free amino groups in the protein with a dicarboxylic anhydride;
- treating a sample of polypeptides with periodate, so that carbohydrates with vicinal cis-diols on glycopeptides will gain a carbonyl functionality;
- labelling this carbonyl functionality with a hydrazide activated biotin;
- digesting the protein sample with a sequence specific endoprotease;
- capturing biotinylated peptides onto an avidin derivitised solid support; and
- detecting the isolated peptides by LC-MS or LC-MS/MS.

The protein sample may be digested with the sequence specific endoprotease before or after reaction of the sample with the hydrazide biotin.

Isolation of Phosphopeptides

Phosphorylation is a ubiquitous reversible post-translational modification that appears in the majority of signalling pathways of almost all organisms as phosphorylation is widely used as a transient signal to mediate changes in the state of individual proteins. It is an important area of research and tools which allow the analysis of the dynamics of phosphorylation are essential to a full understanding of how cells responds to stimuli, which includes the responses of cells to drugs.

Techniques for the analysis of phosphoserine and phosphothreonine containing peptides are well known. One class of such methods is based on a well known reaction for beta-elimination of phosphates. This reaction results in phosphoserine and phosphothreonine

forming dehydroalanine and methyldehydroalanine, both of which are Michael acceptors and will react with thiols. This has been used to introduce hydrophobic groups for affinity chromatography (See for example Holmes C.F., FEBS Lett. **215**(1): 21-24, "A new method for the selective isolation of phosphoserine-containing peptides." 1987). Dithiol linkers have also been used to introduce fluorescein and biotin into phosphoserine and phosphothreonine containing peptides (Fadden P, Haystead TA, Anal Biochem **225**(1): 81-8, "Quantitative and selective fluorophore labelling of phosphoserine on peptides and proteins: characterization at the attomole level by capillary electrophoresis and laser-induced fluorescence." 1995; Yoshida O. *et al.*, Nature Biotech **19**: 379 – 382, "Enrichment analysis of phosphorylated proteins as a tool for probing the phosphoproteome", 2001).

An improved protocol for the beta-elimination based labelling procedure is known. This improved procedure involves barium catalysis. (Byford M.F., Biochem J. **280**: 261-261, "Rapid and selective modification of phosphoserine residues catalysed by Ba2+ ions for their detection during peptide microsequencing." 1991) This catalysis makes the reaction 20-fold faster reducing side-reactions to undetectable levels. A thiol-derivatised biotin reagent could be easily coupled to dehydroalanine or methyldehydroalanine generated from beta-elimination of phosphates using barium catalysis. Thus in a further embodiment of the second aspect of this invention, peptides phosphorylated at serine and threonine may be analysed in a method comprising the steps of:

- pre-cleavage of the polypeptides with cyanogen bromide;
- treating a sample of polypeptides with barium hydroxide to beta-eliminate phosphate groups from phosphoserine and phosphothreonine;
- labelling the resultant dehydroalanine or methyldehydroalanine functionalities with the thiol activated peptide mass tag linked to biotin;
- digesting the protein sample with a sequence specific endoprotease;
- capturing tagged peptides onto an avidin derivitised solid support; and
- detecting the isolated peptides by LC-MS or LC-MS/MS.

The protein sample may be digested with the sequence specific endoprotease before or after reaction of the sample with the thiol-biotin.

A number of research groups have reported on the production of antibodies, which bind to phosphotyrosine residues in a wide variety of proteins. (see for example A.R. Frackelton *et al.*, Methods Enzymol. **201**: 79-92, "Generation of monoclonal antibodies against phosphotyrosine and their use for affinity purification of phosphotyrosine-containing proteins.", 1991 and other articles in this issue of Methods Enzymol.). This means that a significant proportion of proteins that have been post-translationally modified by tyrosine phosphorylation may be isolated by affinity chromatography using these antibodies as the affinity column ligand.

These phosphotyrosine binding antibodies can be used in the context of this invention to isolate terminal peptides from proteins containing phosphotyrosine residues. The tyrosine-phosphorylated proteins in a complex mixture may be isolated using anti-phosphotyrosine antibody affinity columns. In a further embodiment of the second aspect of this invention, a protocol for the analysis of a sample of proteins, which contains proteins phosphorylated at tyrosine, comprises the steps of:

- pre-cleavage of the polypeptides with cyanogen bromide or capping the free amino groups in the protein with a dicarboxylic anhydride;
- treating the sample with a sequence specific cleavage reagent such as Trypsin or Lys-C;
- passing the protein sample through affinity columns contain anti-phosphotyrosine antibodies to isolate only phosphotyrosine modified peptides; and
- detecting the isolated peptides by LC-MS or LC-MS/MS.

Immobilised Metal Affinity Chromatography (IMAC) represents a further technique for the isolation of phosphoproteins and phosphopeptides. Phosphates adhere to resins comprising trivalent metal ions particularly to Gallium(III) ions (Posewitch, M.C. and Tempst, P., Anal. Chem., 71: 2883-2892, "Immobilized Gallium (III) Affinity Chromatography of Phosphopeptides", 1999). This technique is advantageous as it can

isolate both serine/threonine phosphorylated and tyrosine phosphorylated peptides and proteins simultaneously.

IMAC can therefore also be used in the context of this invention for the analysis of samples of phosphorylated proteins. In a further embodiment of the second aspect of this invention, a protocol for the analysis of a sample of proteins, which contains phosphorylated proteins, comprises the steps of:

- pre-cleavage of the polypeptides with cyanogen bromide;
- treating the sample with a sequence specific cleavage reagent such as Trypsin or Lys-C;
- passing the protein sample through an affinity column comprising immobilised metal ions to isolate only phosphorylated peptides; and
- analysing the isolated peptides by LC-MS or LC-MS/MS.

Note that the use of dicarboxylic anhydrides as a solubilisation method is not preferred for use with this particular peptide isolation protocol as the carboxylic acid groups that result from the reaction of these reagents with lysine will have a modest affinity for the IMAC chromatography columns, reducing the specificity of this technique, unless the reagent is removed prior to the IMAC affinity chromatography.

Note that in all of the above procedures, the dicarboxylic anhydrides can be removed immediately prior to the analysis by mass spectrometry.

Analysis of peptides by mass spectrometry

The essential features of a mass spectrometer are as follows:

Inlet System -> Ion Source -> Mass Analyser -> Ion Detector -> Data Capture System

There are preferred inlet systems, ion sources and mass analysers for the purposes of analysing peptides.

Inlet Systems and peptide separations

In the first aspect of this invention an optional chromatographic or electrophoretic separation is used to reduce the complexity of the sample prior to analysis by mass spectrometry. A variety of mass spectrometry techniques are compatible with separation technologies particularly capillary zone electrophoresis and High Performance Liquid Chromatography (HPLC). The choice of ionisation source is limited to some extent if a separation is required as ionisation techniques such as MALDI and FAB (discussed below) which ablate material from a solid surface are less suited to chromatographic separations. For practical purposes, it has been quite costly to link a chromatographic separation in-line with mass spectrometric analysis by one of these techniques. In contrast, dynamic FAB and ionisation techniques based on spraying such as electrospray, thermospray and APCI are all readily compatible with in-line chromatographic separations and equipment to perform such liquid chromatography mass spectrometry analysis is commercially available.

Ionisation techniques

For many biological mass spectrometry applications so called 'soft' ionisation techniques are used. These allow large molecules such as proteins and nucleic acids to be ionised essentially intact. The liquid phase techniques allow large biomolecules to enter the mass spectrometer in solutions with mild pH and at low concentrations. A number of techniques are appropriate for use with this invention including but not limited to Electrospray Ionisation Mass Spectrometry (ESI-MS), Fast Atom Bombardment (FAB), Matrix Assisted Laser Desorption Ionisation Mass Spectrometry (MALDI MS) and Atmospheric Pressure Chemical Ionisation Mass Spectrometry (APCI-MS).

Electrospray Ionisation

Electrospray ionisation requires that the dilute solution of the analyte molecule is 'atomised' into the spectrometer, i.e. injected as a fine spray. The solution is, for example, sprayed from the tip of a charged needle in a stream of dry nitrogen and an electrostatic

field. The mechanism of ionisation is not fully understood but is thought to work broadly as follows. In a stream of nitrogen the solvent is evaporated. With a small droplet, this results in concentration of the analyte molecule. Given that most biomolecules have a net charge this increases the electrostatic repulsion of the dissolved molecule. As evaporation continues this repulsion ultimately becomes greater than the surface tension of the droplet and the droplet disintegrates into smaller droplets. This process is sometimes referred to as a 'Coulombic explosion'. The electrostatic field helps to further overcome the surface tension of the droplets and assists in the spraying process. The evaporation continues from the smaller droplets which, in turn, explode iteratively until essentially the biomolecules are in the vapour phase, as is all the solvent. This technique is of particular importance in the use of mass labels in that the technique imparts a relatively small amount of energy to ions in the ionisation process and the energy distribution within a population tends to fall in a narrower range when compared with other techniques. The ions are accelerated out of the ionisation chamber by the use of electric fields that are set up by appropriately positioned electrodes. The polarity of the fields may be altered to extract either negative or positive ions. The potential difference between these electrodes determines whether positive or negative ions pass into the mass analyser and also the kinetic energy with which these ions enter the mass spectrometer. This is of significance when considering fragmentation of ions in the mass spectrometer. The more energy imparted to a population of ions the more likely it is that fragmentation will occur through collision of analyte molecules with the bath gas present in the source. By adjusting the electric field used to accelerate ions from the ionisation chamber it is possible to control the fragmentation of ions. This is advantageous when fragmentation of ions is to be used as a means of removing tags from a labelled biomolecule. Electrospray ionisation is particularly advantageous as it can be used in-line with liquid chromatography, referred to as Liquid Chromatography Mass Spectrometry (LC-MS).

Matrix Assisted Laser Desorption Ionisation (MALDI)

MALDI requires that the biomolecule solution be embedded in a large molar excess of a photo-excitable 'matrix'. The application of laser light of the appropriate frequency results in the excitation of the matrix which in turn leads to rapid evaporation of the

matrix along with its entrapped biomolecule. Proton transfer from the acidic matrix to the biomolecule gives rise to protonated forms of the biomolecule which can be detected by positive ion mass spectrometry, particularly by Time-Of-Flight (TOF) mass spectrometry. Negative ion mass spectrometry is also possible by MALDI TOF. This technique imparts a significant quantity of translational energy to ions, but tends not to induce excessive fragmentation despite this. Accelerating voltages can again be used to control fragmentation with this technique though.

Fast Atom Bombardment

Fast Atom Bombardment has come to describe a number of techniques for vaporising and ionising relatively involatile molecules. The essential principal of these techniques is that samples are desorbed from surfaces by collision of the sample with accelerated atoms or ions, usually xenon atoms or caesium ions. The samples may be coated onto a solid surface as for MALDI but without the requirement of complex matrices. These techniques are also compatible with liquid phase inlet systems - the liquid eluting from a capillary electrophoresis inlet or a high pressure liquid chromatography system pass through a frit, essentially coating the surface of the frit with analyte solution which can be ionised from the frit surface by atom bombardment.

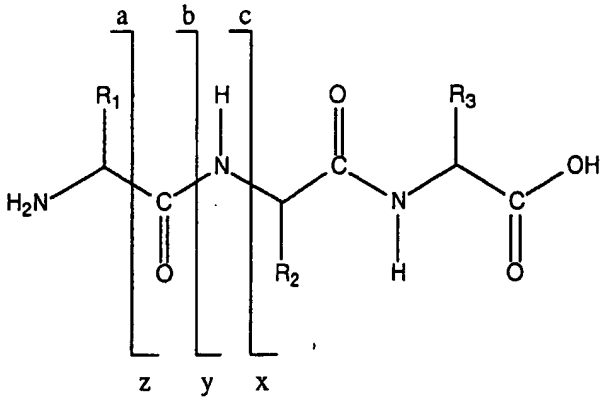
Mass Analysers

Fragmentation of peptides by collision induced dissociation is used in this invention to identify tags on proteins. Various mass analyser geometries may be used to fragment peptides and to determine the mass of the fragments.

MS/MS and MS^n analysis of peptides

Tandem mass spectrometers allow ions with a pre-determined mass-to-charge ratio to be selected and fragmented by collision induced dissociation (CID). The fragments can then be detected providing structural information about the selected ion. When peptides are analysed by CID in a tandem mass spectrometer, characteristic cleavage patterns are observed, which allow the sequence of the peptide to be determined. Natural peptides typically fragment randomly at the amide bonds of the peptide backbone to give series of

ions that are characteristic of the peptide. CID fragment series are denoted a_n, b_n, c_n, etc. for cleavage at the n^{th} peptide bond where the charge of the ion is retained on the N-terminal fragment of the ion. Similarly, fragment series are denoted x_n, y_n, z_n, etc. where the charge is retained on the C-terminal fragment of the ion.



Trypsin and LysC are favoured cleavage agents for tandem mass spectrometry as they produce peptides with basic groups at both ends of the molecule, i.e. the alpha-amino group at the N-terminus and lysine or arginine side-chains at the C-terminus. This favours the formation of doubly charged ions, in which the charged centres are at opposite termini of the molecule. These doubly charged ions produce both C-terminal and N-terminal ion series after CID. This assists in determining the sequence of the peptide. Generally speaking only one or two of the possible ion series are observed in the CID spectra of a given peptide. In low-energy collisions typical of quadrupole based instruments the b-series of N-terminal fragments or the y-series of C-terminal fragments predominate. If doubly charged ions are analysed then both series are often detected. In general, the y-series ions predominate over the b-series.

A typical tandem mass spectrometer geometry is a triple quadrupole which comprises two quadrupole mass analysers separated by a collision chamber, also a quadrupole. This collision quadrupole acts as an ion guide between the two mass analyser quadrupoles. A gas can be introduced into the collision quadrupole to allow collision with the ion stream from the first mass analyser. The first mass analyser selects ions on the basis of their

mass/charge ration which pass through the collision cell where they fragment. The fragment ions are separated and detected in the third quadrupole. Induced cleavage can be performed in geometries other than tandem analysers. Ion traps mass spectrometers can promote fragmentation through introduction of a gas into the trap itself with which trapped ions will collide. Ion traps generally contain a bath gas, such as helium but addition of neon for example, promotes fragmentation. Similarly photon induced fragmentation could be applied to trapped ions. Another favourable geometry is a Quadrupole/Orthogonal Time of Flight tandem instrument where the high scanning rate of a quadrupole is coupled to the greater sensitivity of a reflectron TOF mass analyser to identify the products of fragmentation.

Conventional 'sector' instruments are another common geometry used in tandem mass spectrometry. A sector mass analyser comprises two separate 'sectors', an electric sector which focuses an ion beam leaving a source into a stream of ions with the same kinetic energy using electric fields. The magnetic sector separates the ions on the basis of their mass to generate a spectrum at a detector. For tandem mass spectrometry a two sector mass analyser of this kind can be used where the electric sector provide the first mass analyser stage, the magnetic sector provides the second mass analyser, with a collision cell placed between the two sectors. Two complete sector mass analysers separated by a collision cell can also be used for analysis of mass tagged peptides.

Ion Traps

Ion Trap mass analysers are related to the quadrupole mass analysers. The ion trap generally has a 3 electrode construction - a cylindrical electrode with 'cap' electrodes at each end forming a cavity. A sinusoidal radio frequency potential is applied to the cylindrical electrode while the cap electrodes are biased with DC or AC potentials. Ions injected into the cavity are constrained to a stable circular trajectory by the oscillating electric field of the cylindrical electrode. However, for a given amplitude of the oscillating potential, certain ions will have an unstable trajectory and will be ejected from the trap. A sample of ions injected into the trap can be sequentially ejected from the trap

according to their mass/charge ratio by altering the oscillating radio frequency potential. The ejected ions can then be detected allowing a mass spectrum to be produced.

Ion traps are generally operated with a small quantity of a 'bath gas', such as helium, present in the ion trap cavity. This increases both the resolution and the sensitivity of the device as the ions entering the trap are essentially cooled to the ambient temperature of the bath gas through collision with the bath gas. Collisions both increase ionisation when a sample is introduced into the trap and dampen the amplitude and velocity of ion trajectories keeping them nearer the centre of the trap. This means that when the oscillating potential is changed, ions whose trajectories become unstable gain energy more rapidly, relative to the damped circulating ions and exit the trap in a tighter bunch giving a narrower larger peaks.

Ion traps can mimic tandem mass spectrometer geometries, in fact they can mimic multiple mass spectrometer geometries allowing complex analyses of trapped ions. A single mass species from a sample can be retained in a trap, i.e. all other species can be ejected and then the retained species can be carefully excited by super-imposing a second oscillating frequency on the first. The excited ions will then collide with the bath gas and will fragment if sufficiently excited. The fragments can then be analysed further. It is possible to retain a fragment ion for further analysis by ejecting other ions and then exciting the fragment ion to fragment. This process can be repeated for as long as sufficient sample exists to permit further analysis. It should be noted that these instruments generally retain a high proportion of fragment ions after induced fragmentation. These instruments and FTICR mass spectrometers (discussed below) represent a form of temporally resolved tandem mass spectrometry rather than spatially resolved tandem mass spectrometry which is found in linear mass spectrometers.

Fourier Transform Ion Cyclotron Resonance Mass Spectrometry (FTICR MS)

FTICR mass spectrometry has similar features to ion traps in that a sample of ions is retained within a cavity but in FTICR MS the ions are trapped in a high vacuum chamber by crossed electric and magnetic fields. The electric field is generated by a pair of plate

electrodes that form two sides of a box. The box is contained in the field of a superconducting magnet which in conjunction with the two plates, the trapping plates, constrain injected ions to a circular trajectory between the trapping plates, perpendicular to the applied magnetic field. The ions are excited to larger orbits by applying a radio-frequency pulse to two 'transmitter plates' which form two further opposing sides of the box. The cycloidal motion of the ions generate corresponding electric fields in the remaining two opposing sides of the box which comprise the 'receiver plates'. The excitation pulses excite ions to larger orbits which decay as the coherent motions of the ions is lost through collisions. The corresponding signals detected by the receiver plates are converted to a mass spectrum by Fourier Transform (FT) analysis.

For induced fragmentation experiments these instruments can perform in a similar manner to an ion trap - all ions except a single species of interest can be ejected from the trap. A collision gas can be introduced into the trap and fragmentation can be induced. The fragment ions can be subsequently analysed. Generally fragmentation products and bath gas combine to give poor resolution if analysed by FT analysis of signals detected by the 'receiver plates', however the fragment ions can be ejected from the cavity and analysed in a tandem configuration with a quadrupole, for example.

Prediction of sample peptides

The second aspect of this invention provides a method of predicting, from a list of known polypeptide sequences, the expected products of applying a combination of the solubilisation step followed by a peptide sampling step on the solubilised polypeptides. Figure 1 shows a flow-chart outlining the steps in this algorithm, with a single short example polypeptide. Such an algorithm could be easily implemented as a program in a programming language suitable for the analysis of strings, such as PERL ((Wall, Christiansen *et al.* 1996)). The input to such a program would be a list of known protein sequences. Databases of such sequences are publicly available. Human sequences, for example can be obtained from world wide web servers maintained by the European Bioinformatics Institute (O'Donovan, Martin *et al.* 2002).

The program first simulates the treatment of each polypeptide sequence with the solubilisation reagent. In the example in Figure 1, Cyanogen Bromide is used which cleaves at methionine. This produces a series of cleavage peptides as shown. Amino acids that are expected to be modified by the processes applied to the parent protein are shown in bold capitals. In the cleavage step with Cyanogen bromide, methionine residues are converted to homoserine residues hence they are shown in bold capitals. Note also that in the example, the parent sequence is shown starting with methionine. In many higher eukaryotes, methionine at the N-terminus of a protein is expected to be modified (Dalboge, Bayne *et al.* 1990; Moerschell, Hosokawa *et al.* 1990). Methionine is typically removed if the second amino acid in the sequence has a small radius of gyration i.e. glycine, alanine, serine, cysteine, threonine, proline, and valine. In this example cyanogen bromide would also remove this amino acid. In many proteins the N-terminus is also acetylated and the acetylated residue is often serine, methionine or alanine (Persson, Flinta *et al.* 1985). In this situation multiple predicted entries for N-terminal sample peptides should be included to cover all the possible variants that might be expected.

After the simulated solubilisation of each known polypeptide, the program simulates a sampling process to predict the expected sample peptides and the corresponding modifications of amino acids if they take place. In many processes cysteine is modified. In the ICAT peptide sampling procedure, for example cysteine is modified with biotin. In the example shown in Figure 1 it has been assumed that cysteine any cysteine disulphide bridges have been reduced and any free thiols have been blocked. Iodoacetamide is typically used for this purpose. The mass modification of cysteine is thus marked. In the example shown it has been assumed that the sampling process isolates the N-terminal fragment from each of the solubilisation products according to the disclosure in WO 98/32876. This sampling process relies on blocking the free alpha-amino groups and epsilon amino groups in products of the solubilisation process, these modifications take place at lysine and at free alpha amino groups and these modified amino acids are shown in bold capitals. The blocked polypeptides are then cleaved with a second sequence specific cleavage reagent. In the example in Figure 1 this is trypsin, which cleaves only at

arginine, if the lysine amino groups are blocked. The cleavage process exposes alpha amino groups in non-N-terminal fragments which can then be captured either by reaction with NHS-biotin or an amine-reactive solid support. Thus this sampling process isolates tryptic peptides that have blocked alpha amino groups. The masses of the isolated peptides can then be determined by summing the expected residue masses for each amino acid in the peptide sequence taking into account the expected modifications. Finally the sample peptides are written out to a file with their parent protein in a format such that they are associated in some way.

The algorithm for predicting the masses of peptides sampled from polypeptides solubilised according to the methods of this invention can be implemented in two ways. Specific programs with the parameters for the solubilisation process and sampling process may be pre-specified in the code. Alternatively a general program can be implemented in which the operator is prompted to enter the required parameters. Again considering Figure 1, in the first step of a general algorithm would prompt the user to provide a file with a list of known polypeptide sequences. In the second step, the generalised algorithm would prompt the user to specify the solubilisation process, either CNBr cleavage which results in cleavage at methionine and conversion of methionine to homoserine or reaction of free amino groups with a carboxylic dianhydride, in which case the user would be asked to specify the mass modification that would result at a free amino group, at the alpha amino group or at lysine. A further parameter that the user would be asked to provide is whether the dianhydrides are removed prior to mass spectrometry. In the third step of the process shown in Figure 1, the sampling process parameters require the user to specify cleavage sites at which the sequence specific cleavage reaction takes place, which amino acids have mass modifications and finally the common feature that sampled peptides must share. The specification of the cleavage site and the common feature must be in terms of amino acid and/or sequences of amino acids and these can be specified as regular expressions in the PERL programming (Wall, Christiansen *et al.* 1996), for example in PERL the cleavage of trypsin is defined by the regular expression (K(?!P)|R(?!P)) which specifies cleavage at K or R (lysine and arginine respectively) but not if they are adjacent to P (proline) where the polypeptide sequence is represented by

standard single letter codes. Similarly, the sampling feature should be entered as a regular expression. In simple cases like ICAT (Gygi, Rist *et al.* 1999), the regular expression would simply require matching at least one cysteine residue in each acceptable sample peptide generated by trypsin cleavage. In sampling processes for post-translational modifications, expected sequence motifs at which post-translational modifications can take place would be entered as regular expressions to find matching peptides.

Identification of proteins by matching experimental sample peptide data to predicted data in a database

In the third aspect of this invention the output of the predictive algorithm is stored on a computer readable storage medium such that the predicted sampled peptides can be correlated to their parent peptides. The computer readable storage medium could comprise a relational database in which the data generated by the sampled peptide prediction algorithm is stored in such a fashion that sequences or masses of the peptides can be correlated to their parent polypeptides, see for example Figure 2 in which a pair of example proteins from yeast have been selected and displayed in the table entitled "Parent Polypeptide Sequence Table". The sequences of those polypeptides are stored in a table linked to a key to identify them, a number in this example. Additional data could be stored in this table as well if desired. In the example in Figure 2, the sequences and masses of some sample peptides corresponding to solubilisation of these proteins with cyanogen bromide followed by isolation of the N-terminal fragments of the cyanogen bromide fragments have been predicted with appropriate mass modifications to account for reaction of cysteine with iodoacetamide, conversion of methionine to homoserine and reaction of the alpha and epsilon amino groups with an active ester reagent. The sequences and masses have been stored in the table entitled "Predicted Sampled Peptide Table" with a unique key for each 'sampled' peptide. Additional data could be stored in this table, such as predicted mass-to-charge ratios of the fragmentation products of collision induced dissociation of the peptides to allow direct searching of the database with raw fragmentation data (Yates, Eng *et al.* 1995). The peptides corresponding to a particular mass or sequence are correlated to their parent peptides by their masses or their sequences in corresponding correlation tables. Note that some masses are not uniquely

resolved at any given mass accuracy. In Figure 2, it can be seen that two of the sampled peptides match two different proteins in the mass correlation table assuming a mass accuracy of 15 parts per million (ppm) while in the sequence correlation table all of the peptides have unique sequences linking them to a single parent polypeptide. There will be some examples of peptides, particularly short peptides, where the same sequence may occur in more than one protein but this is a less common occurrence than it is for the mass uniqueness data.

The data in the above format on a computer readable medium can be used, according to the fourth aspect of the invention in a computer aided method to identify proteins. If the methods of the first aspect of the invention are applied to a sample of polypeptides, the end result is a series of sample peptide masses or a series of mass spectrometrically determined fragmentation data for each sample peptide. This experimentally determined data can be used to search a database of predicted sample peptide data generated according to the second aspect of the invention to find predicted sample peptides with predicted masses or predicted fragmentation patterns that match most closely the experimentally determined data. The best matching predicted sample peptides can then be correlated with the corresponding parent polypeptides to find the predicted polypeptides(s) that best match the data thus providing an identification or list of possible identifications for the experimental data.

References

Cardenas, M. S., E. van der Heeft, *et al.* (1997). "On-line derivatization of peptides for improved sequence analysis by micro-column liquid chromatography coupled with electrospray ionization-tandem mass spectrometry." Rapid Commun Mass Spectrom. 11(12): 1271-8.

Crimmins, D. L., D. W. McCourt, *et al.* (1990). "In situ chemical cleavage of proteins immobilized to glass-fiber and polyvinylidenedifluoride membranes: cleavage at tryptophan residues with 2-(2'-nitrophenylsulfenyl)-3-methyl-3'-bromoindolenine to obtain internal amino acid sequence." Anal Biochem **187**(1): 27-38.

Dalboge, H., S. Bayne, *et al.* (1990). "In vivo processing of N-terminal methionine in E. coli." FEBS Lett **266**(1-2): 1-3.

Fontana, A., D. Dalzoppo, *et al.* (1981). "Chemical cleavage of tryptophanyl and tyrosyl peptide bonds via oxidative halogenation mediated by o-iodosobenzoic acid." Biochemistry **20**(24): 6997-7004.

Fontana, A., D. Dalzoppo, *et al.* (1983). "Cleavage at tryptophan with o-iodosobenzoic acid." Methods Enzymol **91**: 311-8.

Gygi, S. P., B. Rist, *et al.* (1999). "Quantitative analysis of complex protein mixtures using isotope-coded affinity tags." Nat Biotechnol. **17**(10): 994-9.

Kamo, M. and A. Tsugita (1998). "Specific cleavage of amino side chains of serine and threonine in peptides and proteins with S-ethyltrifluorothioacetate vapor." Eur. J Biochem **255**(1): 162-71.

Kaplan, H. and Oda (1983). " Selective isolation of free and blocked amino-terminal peptides from enzymatic digestion of proteins." Anal Biochem **132**(2): 384-388.

Mahoney, W. C. and M. A. Hermodson (1979). "High-yield cleavage of tryptophanyl peptide bonds by o-iodosobenzoic acid." Biochemistry **18**(17): 3810-4.

Moerschell, R. P., Y. Hosokawa, *et al.* (1990). "The specificities of yeast methionine aminopeptidase and acetylation of amino-terminal methionine in vivo. Processing of altered iso-1-cytochromes c created by oligonucleotide transformation." J Biol. Chem. **265**(32): 19638-43.

Nieto, M. A. and Palacian (1983). " Effects of temperature and pH on the regeneration of the amino groups of ovalbumin after modification with citraconic and dimethylmaleic anhydrides." Biochim Biophys Acta **749**(2): 204-210.

O'Donovan, C., M. J. Martin, *et al.* (2002). "High-quality protein knowledge resource: SWISS-PROT and TrEMBL." Brief Bioinform. **3**(3): 275-84.

Palacian, E., P. J. Gonzalez, *et al.* (1990). " Dicarboxylic acid anhydrides as dissociating agents of protein-containing structures." Mol Cell Biochem **97**(2): 101-111.

Persson, B., C. Flinta, *et al.* (1985). "Structures of N-terminally acetylated proteins." Eur J Biochem **152**(3): 523-7.

Roth, K. D., Z. H. Huang, *et al.* (1998). "Charge derivatization of peptides for analysis by mass spectrometry." Mass Spectrom Rev **17**(4): 255-74.

Smith, B. J. (1994). "Chemical cleavage of proteins." Methods Mol Biol. **32**: 297-309.

Smith, B. J. (1997). "Chemical cleavage of polypeptides." Methods Mol Biol. **64**: 57-72.

Tsugita, A., M. Kamo, *et al.* (1998). "Additional possible tools for identification of proteins on one- or two-dimensional electrophoresis." Electrophoresis **19**(6): 928-38.

Tsugita, A., K. Takamoto, *et al.* (1992). "C-terminal sequencing of protein. A novel partial acid hydrolysis and analysis by mass spectrometry." Eur J Biochem **206**(3): 691-6.

Vestling, M. M., M. A. Kelly, *et al.* (1994). "Optimization by mass spectrometry of a tryptophan-specific protein cleavage reaction." Rapid Commun Mass Spectrom **8**(9): 786-90.

Wall, L., T. Christiansen, *et al.* (1996). Programming Perl, O'Reilly & Associates, Inc.

Wu, J. and J. T. Watson (1998). "Optimization of the cleavage reaction for cyanylated cysteinyl proteins for efficient and simplified mass mapping." Anal Biochem **258**(2): 268-76.

Yates, J. R., J. K. Eng, *et al.* (1995). "Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database." Anal Chem. **67**(8): 1426-1436.

CLAIMS:

1. A method of solubilising a polypeptide, which polypeptide is insoluble or sparingly soluble in an aqueous medium, which method comprises either:

(a) contacting the polypeptide with a sequence specific cleavage agent in a non-aqueous medium, to produce a product which is soluble in an aqueous medium; or

(b) contacting the polypeptide with a dicarboxylic anhydride to modify one or more amino groups in the polypeptide, to produce a product which is soluble in an aqueous medium.

2. A method according to claim 1, wherein the sequence specific cleavage agent comprises cyanogen bromide, BNPS-skatole or iodosobenzoic acid.

3. A method according to claim 1 or claim 2, wherein the sequence specific cleavage agent cleaves at a methionine residue, a tryptophan residue, a cysteine residue, a threonine residue or a serine residue.

4. A method according to claim 1, wherein the dicarboxylic acid caps the one or more amino groups.

5. A method for characterising a polypeptide which is insoluble or sparingly soluble in an aqueous medium, which method comprises:

(a) solubilising the polypeptide according to a method as defined in any of claims 1-4, to form a solubilised product;

(b) optionally cleaving the solubilised product to form one or more peptides;

(c) identifying one or more peptides characteristic of the polypeptide;

(d) characterising the polypeptide on the basis of the one or more identified peptides.

6. A method according to claim 5, wherein the one or more peptides are isolated.

7. A method according to claim5 or claim 6, wherein the one or more peptides characteristic of the polypeptide are identified using mass spectrometry.

8. A method according to any of claims 5-7, wherein dicarboxylic anhydride is used in solubilising the polypeptide, and is removed before identifying the one or more peptides.

9. A method according to any of claims 5-8, wherein the one or more peptides are separated by liquid chromatography prior to identifying the peptides.

10. A method according to any of claims 5-9, wherein the method further comprises comparing the identified peptides with peptides in a database, in which combinations of peptides are relatable to parent polypeptides, to characterise the polypeptide.

11. A method of producing a database for identifying a polypeptide, which method comprises:

(a) selecting one or more parent polypeptides;

(b) calculating one or more peptide sequences that would result from one or more putative reactions that each parent polypeptide could undergo;

(c) storing the calculated peptide sequences in a database such that the sequences that would result for a specific parent polypeptide undergoing its selected reaction are relatable to that parent polypeptide when undergoing that reaction.

12. A method according to claim 11, wherein a plurality of putative reactions are selected for one or more of the parent polypeptides.

13. A method according to claim 11 or claim 12, wherein the putative reactions are selected from solubilising reactions and sequence specific cleavage reactions.

14. A method according to any of claims 11-13, wherein one or more of the putative reactions is a multi-step reaction.

15. A method according to claim 14, wherein a multi-step reaction comprises a first solubilising step and a subsequent sequence specific cleavage step.

16. A method according to any of claims 11-15, which method is carried out using a computer program.

17. A database obtainable by a method as defined in any of claims 11-15.

18. A computer-readable storage medium comprising a database obtainable according to a method as defined in claim 16.

19. A relational database comprising:

(a) a table of parent polypeptide sequences; and

(b) a table of peptides, which peptides are expected products of putative reactions of polypeptides from the table of polypeptides;

wherein each of the parent polypeptide sequences is relatable by the database to one or more peptides from the table of peptides.

20. A kit for characterising a polypeptide, which kit comprises:

(a) a solubilisation agent;

(b) optionally a reagent for modifying a reactive group in a peptide or polypeptide;

(c) a sequence specific cleavage agent; and

(d) a means for selectively isolating a subset of peptides generated by the cleavage agent.

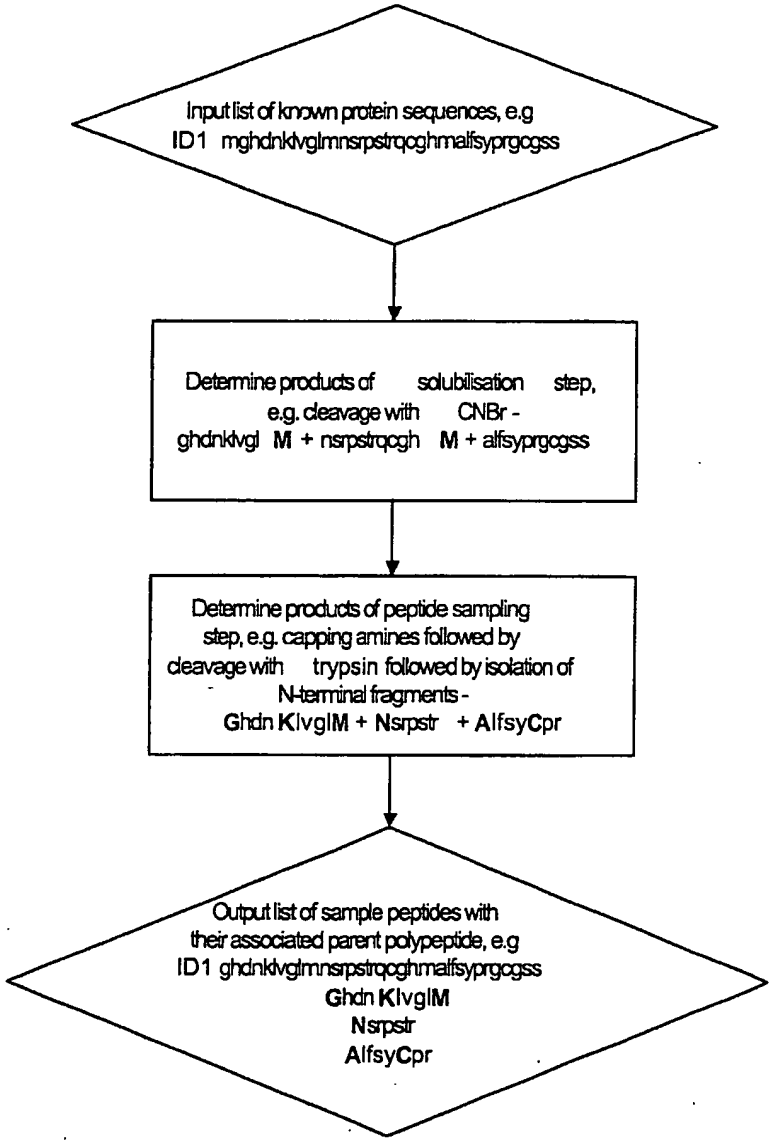

Input list of known protein sequences, e.g
ID1 mghdnklvglmnsrpstrqcghmalfsyprgcgss
Determine products of solubilisation step,
e.g. cleavage with CNBr -
ghdnklvgl M + nsrpstrqcgh M + alfsyprgcgss
Determine products of peptide sampling
step, e.g. capping amines followed by
cleavage with trypsin followed by isolation of
N-terminal fragments -
Ghdn KlvglM + Nsrpstr + AlfsyCpr
Output list of sample peptides with
their associated parent polypeptide, e.g
ID1 ghdnklvglmnsrpstrqcghmalfsyprgcgss
Ghdn KlvglM
Nsrpstr
AlfsyCpr

Figure 1

Parent Polypeptide Sequence Table:

Protein ID	Protein Sequence
5	MVQRWLYSINAKDIAVLYRMLAIFSGMAGTAMSLIIRLELAAPGSQYLHGNSQLRNGAPTSAY ISLMRTALVLWIINRYLKHMTNSVGANFTGIMACHKIPMSVQGVKCYMVRLINFLQVFRIT ISSYHLDMVKQVWLFYVEVIRLWFHVLDSTGSVKKMKDTNNTKGNTKSEGSTERGNSGVDRG MVVPNTQMKMRFLNQVRYYSVNNNLKMGKDTNMELSKDTSTSDLLEFEKLVMDNMNEENM NNNLLSIMKNVDMLMLAYNRIKSKPGNMTRGTTLETLDGMNMMYLNKLSNELGTGKFKFKP MRMVNFKPKDGMRFLSVGNPRDKIVQEVMRMILDTIFDKKMSTHSHGFRKNMSCQTAIWEV RNMFGGSNWFEVDLKKCFDTISHDLIIKELKRYSDKGFRDLVYKLLRAGYIDEKGTYHKPML GLPQGSLISPILCNIVMTLVDNWLEDYNLYNKGKVKKQHPTYKKLSRMIAKAKMFSTRLKLH KERAKGPTHYNDFNFKRMKYVRYADDLIGVLGSKNDCKMIKRDLNNFLNSLGLTMNEEKTL ITCATEIPARFLGYNSITPLKRMPTVTKCTRGKTIRSRNTTRHINAPRDLINKLATNGYCKHNK NGRMGVPTRVGRWTYEEFRTENNYKALGRGILNYYKLATNYKRLRERTYYVLYYSCVLTLAS KYRLKIMSKTIKKFGYNLNIENDKLIANFPRNTHDNKKIENHGMFMYMSEAKVTDPFEYIDSI KYMLPTAKANFNKFCSICNSTIDVEMHHVKQLHRGMLKATKDYITGRMITMNRKQPLCKQC HIKTHKNKFKNMGFGM
62	MSAPTMRSTSILTEHLGYPFSLVDDINAVNEMYKCTAAMEKYLLSKSKIGEEDYGEEIKSGV AKLESLLENSVDKNFDKLELYVLRNVLRIPEEYLDANVFRLENQKDLVIVDENELKKSEEKLR EKVNDVELAFKKNEMLLKRVTKVKRLLFTIRGKQKLNELLKCKDDVQLQKILESLKFDDTM TLLTDSLRKLYVDSESTSSTEEVEALLQRLKTNGKQNNKDFRTRYIDDRTNNVLRKLGLLGDKE DEKQSAKPDARTQAGDIVSDIEEFQLDLLDDVL

Sequence Correlation Table

Peptide ID	Protein ID
3213	5
3215	5
3231	5
4561	62
4562	62

Mass Correlation Table at 15 ppm:

Peptide ID	Protein ID
3213	5
3213	62
3215	5
3231	5
4561	62
4561	5
4562	62

Predicted Sampled Peptide Table:

Peptide ID	Peptide Sequence	Peptide Mass
3213	ACHKTPM	1094.866967
3215	VKQVWLFYVEVIR	1981.266267
3231	SEAKVTDPFEYIDSIKYM	2529.504294
4561	YKCTAAM	1094.856967
4562	EKYLLSKSKIGEEDYGEEIKSGVAKLE SLLENSVDKNFDKLELYVLR	6549.748229

Figure 2